Free-Energy Minimization and the Dark-Room Problem

frontiers in
PSYCHOLOGY

HYPOTHESIS AND THEORY ARTICLE
published: 08 May 2012
doi: 10.3389/fpsyg.2012.00130

Free-energy minimization and the dark-room problem

Karl Friston*[1]*, Christopher Thornton*[2] *and Andy Clark*[3]

[1] *The Wellcome Trust Centre for Neuroimaging, University College London, London, UK*
[2] *Informatics, University of Sussex, Brighton, UK*
[3] *School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh, UK*

Edited by:
Lars Muckli, University of Glasgow, UK

Reviewed by:
Anil K. Seth, University of Sussex, UK
Philippe G. Schyns, University of Glasgow, UK
Jakob Hohwy, Monash University, Australia

***Correspondence:**
Andy Clark, School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh EH8 9AD, UK.
e-mail: andy.clark@ed.ac.uk

Recent years have seen the emergence of an important new fundamental theory of brain function. This theory brings information-theoretic, Bayesian, neuroscientific, and machine learning approaches into a single framework whose overarching principle is the minimization of surprise (or, equivalently, the maximization of expectation). The most comprehensive such treatment is the "free-energy minimization" formulation due to Karl Friston (see e.g., Friston and Stephan, 2007; Friston, 2010a,b – see also Fiorillo, 2010; Thornton, 2010). A recurrent puzzle raised by critics of these models is that biological systems do not seem to avoid surprises. We do not simply seek a dark, unchanging chamber, and stay there. This is the "Dark-Room Problem." Here, we describe the problem and further unpack the issues to which it speaks. Using the same format as the prolog of Eddington's *Space, Time, and Gravitation* (Eddington, 1920) we present our discussion as a conversation between: an information theorist (Thornton), a physicist (Friston), and a philosopher (Clark).

Keywords: free-energy principle, Bayesian brain, surprise, optimality

Philosopher: The "free-energy principle" (see e.g., Friston and Stephan, 2007; Friston, 2010a) suggests that all biological systems are driven to minimize an information-theoretic (not thermodynamic, though the two are mathematically close) quantity known as "free energy." Free energy, as here defined, bounds surprise, conceived as the difference between an organism's predictions about its sensory inputs (embodied in its models of the world) and the sensations it actually encounters. In this discussion, surprise is used explicitly as a measure of improbability from information theory. This is also known as *surprisal* or self information. Although the psychological notion of surprise is distinct, events with high surprisal are generally surprising. Organisms that succeed, the free-energy principle mandates, do so by minimizing their tendency to enter into this special kind of surprising (that is, non-anticipated) state. But at first sight this principle seems bizarre. Animals do not simply find a dark corner and stay there. Play and exploration are core features of many life-forms. Can the free-energy principle really be justified in information-theoretic terms?

Theorist: It seems right to begin with a few general remarks about Information Theory. Remarkably, it is now more than six decades since Claude Shannon set out this framework, with its beautifully simple, core idea of equating generation of information with reduction of uncertainty (i.e., "surprise"). Showing that surprise could be quantified in terms of the range of choice applying, Shannon gave the framework mathematical teeth, enabling the extraordinary diversity of applications that continue to this day.

Part of the excitement about Information Theory has long related to its potential for explaining natural processes and phenomena. One proposition has been that we can explain adaptive behavior by saying it is a mechanism by which agents "reduce their surprise" about the environment they inhabit. This idea appeals in a number of ways. Not only does it promise to provide adaptive behavior with a mathematical foundation, it also generalizes straightforwardly to the case of learning and cognition. The acquisition of any form of knowledge can be viewed in the same way, as an attempt to reduce surprise. Last but not least, there is the intriguing link between informational uncertainty and physical disorder. Mathematically, they are identical. Cutting a long story short, it then becomes possible to envisage a deep, underlying unity connecting generative processes of adaptation, mind, and life.

A proposal of this general form has recently been made in the "free energy" framework, particularly associated with the work of the *Physicist* in the debate (Friston). To the underlying "surprise-reduction" hypothesis, this adds a rich assembly of mathematics, with the aim of explaining many architectural and developmental aspects of brain function. Controversy continues, however, regarding the value of the underlying hypothesis. The proposal that adaptive behavior can be interpreted in broadly informational terms seems beyond dispute. In brief, we can certainly view the process by which agents adapt to their environments as a process by which they reduce their surprise. The problem is we can also view it the other way around, seeing the situation in terms of agents reducing their surprise by adapting to the environment.

Furthermore, the hypothesis seems to have an element missing. Surprise is calculated with reference to the agent's interpretation or "model" of the world. It is the way probabilities are assigned to features of the world that fixes the level of surprise in any given case. Let us say I turn up to watch a football match and discover that the players are all running around the edge of the pitch for some reason. I might respond adaptively, by re-interpreting the situation as an athletics meeting rather than a football match. But this only functions to reduce surprise if it is calculated from probabilities arising from my *original* interpretation. Surprise must be defined relative to the right interpretation (i.e., model of probability) in order to get the adaptive effect. Invoking the surprise-reduction hypothesis as an account of how this interpretation comes into

existence then seems to produce a circular explanation. Some crucial element must be missing.

Initially, it seems we can escape the circularity by taking the position that agents do not form any substantive interpretation. We envisage them attending to uninterpreted sensory data, instead. On this basis, the probabilities that mediate measurement of surprise can be taken to be the frequencies with which different data are acquired. This strategy does not really work, however. If this really is the basis on which agents operate, reduction of surprise dictates blocking out sensory data altogether. Alternatively but equally absurdly, agents should proceed directly to the least stimulating environment and stay there. That is to say, they should take up position in the nearest "dark room" and never move again. This will always be the best way to reduce surprise for an agent that operates in the absence of an adaptively appropriate interpretation. This, then, is the so-called "dark-room problem" that provides the title and focus for the debate.

Appealing as it seems at the outset, then, the surprise-reduction hypothesis leads into a serious tangle. If we allow unlimited rein over the interpretations agents are assumed to apply, the dark-room problem can be eliminated. But the hypothesis then seems to be stating something that is true by definition. If we go the other way, ruling out substantive interpretation, the hypothesis becomes contentful but dictates that agents will tend to behave very stupidly indeed. There seems to be something wrong. All three of us are convinced, however, that there is also something importantly *right* in the general idea.

I would like to start, then, by putting to the Physicist the central question about explanatory content. Specifically, should we view the surprise-reduction hypothesis (of adaptive behavior) as having content *independently*? Or should we take it to be part of some larger explanatory package, in which other elements resolve the problems noted. If it is part of some larger package, what other principles might be involved other than reduction of surprise?

Physicist: From the point of view of the free-energy formulation there is no need to recourse to any other principles. Of course, one might find that one's favorite principle emerges from a particular application of the free-energy principle; however, the whole point of the free-energy principle is to unify all adaptive autopoietic and self-organizing behavior under one simple imperative; *avoid surprises and you will last longer*.

It might be useful to contextualize the free-energy principle in relation to other principles here. From an information theory or statistical perspective, free-energy minimization lies at the heart of variational Bayesian procedures (Hinton and van Camp, 1993) and has been proposed as a *modus operandi* for the brain (Dayan et al., 1995) – a *modus operandi* that appeals to Helmholtz's unconscious inference (Helmholtz, 1866/1962). This leads naturally to the notion of perception as hypothesis testing (Gregory, 1968) and the Bayesian brain (Yuille and Kersten, 2006). Indeed, some specific neurobiological proposals for the computational anatomy of the brain are based on this formulation of perception (Mumford, 1992). Perhaps the most popular incarnation of these schemes is predictive coding (Rao and Ballard, 1999). The free-energy principle simply gathers these ideas together and summarizes their imperative in terms of minimizing free energy (or surprise). However, the free-energy principle brings something else to the table – it says that action should also minimize free energy. With this simple addition, we are now in a position to consider behavior and self organization; however, the same basic principle remains – namely, minimizing free energy or surprise.

Having said this, I think you are right to invoke the notion of a "larger package," in that "surprise" is minimized over multiple scales. For example, the fact that you can reinterpret football as athletics rests upon having an internal model of both football and athletics. The acquisition of these concepts depends upon free-energy minimization during learning. The fact you survive long enough to learn rests on free-energy minimization at an evolutionary scale; and so on.

Avoiding surprises means that one has to model and anticipate a changing and itinerant world. This implies that the models used to quantify surprise must themselves embody itinerant wandering through sensory states (because they have been selected by exposure to an inconstant world): Under the free-energy principle, the agent will become an optimal (if approximate) model of its environment. This is because, mathematically, surprise is also the negative log-evidence for the model entailed by the agent. This means minimizing surprise maximizes the evidence for the agent (model). Put simply, the agent becomes a model of the environment in which it is immersed. This is exactly consistent with the Good Regulator theorem of Conant and Ashby (1970). This theorem, which is central to cybernetics, states that "every Good Regulator of a system must be a model of that system." This means a Dark-Room agent can only exist if there are embodied agents that can survive indefinitely in dark rooms (e.g., caves). In short, Dark-Room agents can only exist if they can exist. The tautology here is deliberate, it appeals to exactly the same tautology in natural selection (Why am I here? – because I have adaptive fitness: Why do I have adaptive fitness? – because I am here). Like adaptive fitness, the free-energy formulation is not a mechanism or magic recipe for life; it is just a characterization of biological systems that exist. In fact, adaptive fitness and (negative) free energy are considered by some to be the same thing.

The particular minimum free-energy solutions associated with that existence will be unique to each conspecific and its econiche. Interestingly, Dark-Room agents do exist: Troglophiles have evolved to model and navigate environments like caves (Barr and Holsinger, 1985). So why do they exist? Surprise is a function of sensations and the agent (model) itself. This means that the surprise can be reduced by changing sensory input (action), predictions of that input (perception), or the model *per se*; through evolution to minimize free energy or maximize free-fitness (Sella and Hirsh, 2005). Evolutionary or neurodevelopmental optimization of a model is distinct from perception and entails changing the form and architecture of an agent. In this sense, every agent represents a viable solution to the free-energy minimization problem that is supported by the real world.

Technically, the resolution of the Dark-Room Problem rests on the fact that average surprise or entropy $\mathbf{H}(s \mid m)$ is a function of sensations *and the agent (model) predicting them*. Conversely, the entropy $\mathbf{H}(s)$ minimized in dark rooms is only a function of sensory information. The distinction is crucial and reflects the fact that surprise only exists in relation to model-based expectations. The free-energy principle says that we harvest sensory signals

that we can predict (cf., emulation theory; Grush, 2004); ensuring we keep to well-trodden paths in the space of all the physical and physiological variables that underwrite our existence. In this sense, every organism (from viruses to vegans) can be regarded as a model of its econiche, which has been optimized to predict and sample from that econiche. Interestingly, free energy is used explicitly for model optimization in statistics (e.g., Yedidia et al., 2005) using exactly the same principles.

This means that a dark room will afford low levels of surprise if, and only if, the agent has been optimized by evolution (or neurodevelopment) to predict and inhabit it. Agents that predict rich stimulating environments will find the "dark room" surprising and will leave at the earliest opportunity. This would be a bit like arriving at the football match and finding the ground empty. Although the ambient sensory signals will have low entropy in the absence of any expectations (model), you will be surprised until you find a rational explanation or a new model (like turning up a day early). Notice that average surprise depends on, and only on, sensations and the model used to explain them. This means an agent can compare the surprise under different models and select the best model; thereby eluding any "circular explanation" for the sensations at hand.

Philosopher: It seems to me there are a number of important distinctions coming out in this. First, we distinguish the various time scales of adaptation. Second, we recognize that the gross bodily form, biomechanics, and gross initial neural architecture of the agent all form part of the (initial) "model" and that this model is further tuned by learning and experience. Third, we distinguish low absolute entropy of sensory signals (where that just amounts, I suppose, to low variability) from lack of surprise relative to the overall model delivered by the process of free-energy minimization.

At this point I would like to introduce another, perhaps not unrelated, issue. In raising it, I follow the advice of Sellars, 1962, p. 37] who famously wrote that the aim of philosophy "is to understand how things in the broadest possible sense of the term hang together in the broadest possible sense of the term." One question that thus arises concerns the relation between all this talk about surprise minimization *in the informational sense*, and surprise minimization from the point of view of an intelligent agent. Perhaps it is too obvious to be worth stating, but we should also bear in mind that the two are remarkably distinct. One way to see this is to reflect that the state of the brain that most thoroughly – across all the time scales of adaptation – minimizes informational surprise may, at times, be a state that corresponds to a very surprising event or percept as far as the agent herself is concerned. For example, if I perceive a pink elephant in the middle of the room, that percept must itself be the one that – taking all those time scales (i.e., experience) into account – most effectively minimizes the long-term average of surprise about such data, conditioned on a model (again, in the inclusive sense of model). We should thus remind ourselves that even surprise relative to our best model can be tolerated, as evidenced by surprisingness to the conscious agent who may often – though not *too* often on pain of death – find herself in quite surprising and unexpected situations.

But despite thus agreeing on very many fronts with *physicist* there is something correct, it seems to me, about the shape of the most general worry that *theorist* raises. What is correct is the worry that the story on offer may, in at least one important sense, turn out to be a little bit *too* accommodating. For the appeals to (the minimization of) free energy and informational surprise are, as we just saw, compatible with a very large space of possible strategies, lifestyles, and cognitive architectures. The whole shape of the space of prior expectations is what determines surprise. That shape is specific to different species and may also vary (as a result of learning) between different individuals. For example there might, as we noted, indeed be species for whom the best way to avoid decay and disorder is to find the dark (but nutrient-rich) place and stay in it. We are not those creatures, since that is not the nature of our lifestyle/environment. Given the needs of our lifestyle, and the uncertainty and variability of the environment, policies that will minimize *our* free energy, and with it an upper bound on informational surprise, will be ones that mandate motion, search, discovery, and constructive action. Now in one way this is clearly unproblematic. The briefest glance at the staggering variety of biological life-forms tells us that whatever fundamental principles are sculpting life and mind, they are compatible with an amazing swathe of morphological, neurological, and ethological outcomes. But in another way it may still seem disappointing, if what we want to understand is the specific and detailed functional architecture of the mature human mind. For the distance between these general principles and the specific solutions to adaptive needs first embodied in young human brains, and then morphed and filtered through learning and experience, seems daunting. This problem is exacerbated when we realize that all that learning and experience is itself conditioned by the work of many previous generations, whose legacy of stacked designer environments means that mature human brains have been nurtured in some very singular cognitive "vats" indeed.

Bearing all that in mind, it seems that we cannot know (*a priori* as it were) just how much of the work in explaining and understanding actual human minds can be done by direct appeal to the free-energy framework, and how much must be done instead by the discovery (by whatever other means cognitive science has available) of the variety of idiosyncratic and evolutionary and developmentally path-dependent ploys, routines, and policies that determine the responses of mature human agents. Contrariwise, the greatest virtue of the free-energy framework, it seems to me, is that it reveals the underlying unity beneath that superficially heterogeneous array of ploys and policies, displaying bodily form, biomechanics, learning, niche-construction, perception, and action as manifestations of a single ongoing adaptive imperative to reduce informational surprise. The resulting unified model of brains, bodies, and active, environmentally embedded agents seems to me to be one of the most exciting new developments in the ancient quest to understand mind and its place in nature.

Physicist: Your (pragmatic) endorsement of the free-energy principle is very consistent with its motivation. And I should probably concur with your gracious conclusions and leave it at that: However, I cannot resist noting your disappointment that deployment of the principle does not prescribe the exact nature of biological agents and in particular the "detailed functional architecture of the mature human mind." I would not be so pessimistic. We have not touched upon the nature of "models" in

this discussion but (to take one example) a key component of any model (especially for social agents) will be a model of conspecifics. In other words, my model of the world will include a model of you, which will include your model of me and so on. Exploring the implications of this (even in thought experiments) leads to some interesting constraints on the "functional architecture" of the mind. For example, modeling your model (of my model of your model . . .) induces an infinite regress, which is bounded by the physics of my brain. This rests on the assertion that an agent is a model of its world. Because the agent is physically embodied, so is its model (Kirsh, 1996; Ziemke, 2002). This means we cannot entertain infinite regresses when modeling interactions with other agents; i.e., our models are bounded (bounded rationality is an important constraint in game theory and economics). A bound on recursive "theory of mind" (models of models of models . . .) becomes quite acute when one considers that the most important conspecific in the world is me. This means my representation of my representation of my representation . . . is bounded at some (low) finite order. In short, I can never conceive of what it is like to be me, because that would require the number of recursions I can physically entertain, plus one. I introduce this partly to amuse you (and partly to emphasize the philosophical naivety of the *physicist*) but mostly to suggest that there are many interesting questions and implications that have yet to be unpacked in this (free energy) context.

Philosopher: It is indeed interesting to see that even superficially "idiosyncratic" solutions of the kind routinely displayed in the literature on bounded rationality, fast-and-frugal heuristics, etc., might in principle reflect direct applications of the free-energy principle to reasoning in highly complex domains (such as the understanding of self and others). Nonetheless, considerable space remains between the (surely correct) assertion that such direct explanations might be available, and the assertion that such explanations will always be available. I do not yet see what would conclusively establish the universal claim. But if the universal claim is not established, then it does indeed seem as if there remains an important empirical question concerning the ultimate scope of the explanations that proceed directly (albeit across multiple timescales) from the free-energy principle to the full swathe of features, strategies, and properties distinctive of human cognition.

It seems to me that the claim that the free-energy principle "prescribes" the nature of biological agents may be heard in two distinct ways, then. In the first (I think incorrect) way, it would mean that everything about the agent here-and-now can be explained by considering the ways in which its structure is such as to minimize (over multiple time scales) free energy in its exchanges with the environment. In the second (I think correct) way it would mean that everything about the agent has been selected under the constraints of free-energy minimization, working alongside a swathe of contingent influences. If this is right, then it is an (interesting and important) empirical question just how much of the human mind and even our adult, "culturally marinated" cognitive architecture is to be explained by direct appeal to the free-energy framework, and how much is determined by those messier processes of historical path-dependence.

Theorist: I am not sure it is the processes of historical path-dependence that are key. I do agree with the idea, though, that it is the inability of the principle to explain the operation of these "contingent influences," while at the same time requiring them to be in full working order, that seems the crux of the problem. If these processes are not clearly specified, we have a truly vast range of situations that will comfortably accommodate the proposed principle.

This matters, because the level of surprise an agent registers regarding specific sensory stimulation all depends on how the entropy is measured. This depends on how probabilities are assigned to the different constituents. Different assumptions about the relevant probabilities then yield different measurements of surprise, and thus different "surprise reducing" strategies.

Consider football again, but now with the idea that we want to construct a set of usefully autonomous cambots. Cambots are adaptive agents that are meant to produce nicely angled video recordings using camera-style sensory stimulation. Let us name these agents MIN1, MIN2, MIN3, and MIN4. All will minimize surprise; but each will be able to choose how probabilities are determined. Consider what happens when one of these agents acquires visual stimulation containing a blob of pink pixels in the top-right corner of its sensory image.

MIN1 works on the basis that the probability of a particular pixel having a particular color is determined by interactions between ambient light and 3D structures in the visual field. Calculation of probabilities for pixels in the pink blob then yields an intermediate level of surprise, reflecting the fact that pink is a possible but not very likely color for a 3D structure in a football stadium. Attempting to reduce this surprise, MIN1 produces a series of physical motions (crashing into a barrier in the process) in an attempt to "crisp up" the visual border of the blob.

MIN2 works on the basis that the pixel-color probabilities are determined by the statistical properties of the image itself. Determination of probabilities on this basis then yields a very low level of surprise, reflecting the fact that, for MIN2, an image comprised exclusively of blobs is by far the least surprising. No action is required; none is taken.

MIN3 works on the basis that pixel-color probabilities are determined solely by local textures based on adjacent pixel values. MIN3 believes fields of (pink) pixels with no texture are improbable and registers a high degree of sensory surprise. Aiming to reduce this, it commences a random exploration of the stadium, eventually finding a back-room containing a tartan flag. Offering a rough approximation of a checkerboard pattern, this produces the stimulation that is minimally surprising within MIN3's pixel-adjacency probability model.

Finally, MIN4 works on the basis that pixel-color probabilities are determined by events and patterns in the real world, pretty much as humans conceive them, including such things as the rules of football, social norms, dynamic properties of 3D objects and so forth. Detecting that the pink blob in question is actually the visual stimulus resulting from a pink elephant standing in the middle of the pitch, MIN4 determines probabilities yielding an extremely high level of surprise. Surprise-reduction kicks in, producing immediate flight, the agent vigorously seeking out a more predictable environment.

The point is, of course, that such examples can be continued indefinitely. If the critical question of how probabilities are determined is left unsettled, there is no limit to the ways we can

envisage agents determining probability distributions, and thus no limit to the range of strategies that might fulfill "surprise reduction" in practice. The reduce-surprise/live-longer hypothesis is thus consistent with a very large number of interpretations of informed, adaptive behavior.

Philosopher: I am not sure the issue is quite so clear-cut. Theorist's worry is similar in some ways to my own worry that a great deal of explanatory work remains to be done even once the free-energy framework is in place. But recall that my worry was that the precise ways in which we think and behave may reflect a process of evolutionary and cultural search that is sensitive to many factors that are arbitrary (perhaps because highly path-dependent) with respect to the free-energy model. Notice that this is meant to locate a genuine shortfall, not merely to reflect the need to minimize informational surprise across multiple time scales. By contrast, it seems to me that the last example that *Theorist* gives (MIN1 to MIN4) invites a timescale-based response. Assuming that the bots will be subject to some kind of selection pressure, so that only bots that act in certain ways get to survive and reproduce, then the differences in how they compute probabilities and measure surprise will in the end be responsive to these needs. Given a bot-niche, surprise minimization will need to invoke a suitable measure of surprise. That means that what *Theorist* depicts as an unconstrained choice of measures of surprise is in fact a highly constrained choice.

Physicist: Perhaps a useful way to think about the multitude of different surprises agents or robots might entertain; and how they are constrained by free energy is as follows: different surprises rest on different assumptions about the world that can be cast as prior beliefs and are therefore part of the model. This means there is a unique surprise for every model but the strategy is the same – reduce surprise. These prior beliefs could be implicit in the form of the agent (e.g., the wavelength selectivity of photoreceptors reflects prior assumptions about the wavelength of ambient light) or its parameters (e.g., neuronal connection strengths optimized by experience-dependent plasticity). Crucially, these priors (assumptions) are updated and optimized to minimize free energy because they are part of the model.

The key point Theorist makes with the MIN "cambots" series is that there can be an infinite number of models ("cambots"). However, only one will have the lowest (average) free energy. This is the "cambot" that would be selected in an evolutionary setting or by a "cambot" designer (cf., Buason et al., 2005). This selection is just another instance of free-energy optimization but operating at the level of models. At the same time, each model is trying minimize its own free energy.

Philosopher: This is the part I still do not get. I can see that, if we assume a bodily form, a neural architecture, and an environmental niche, then it is likely that one cambot will, over any finite interval, have the lowest average free energy. I think I can see, too, that even if we allow bodily form and neural architecture to vary amongst individuals, there will still be a winner in the competition to be the most efficient inhabitant of the niche (hence the bot with the lowest average free energy). But we do have to assume a niche (a job relative to which the bots are assessed) and fix a timescale (an interval over which to average). As far as I can see, nothing in the free-energy story fixes these. Does that matter?

Physicist: Free-energy minimization automatically takes care of this and operates at multiple scales. This is because the average free energy (over time) is minimized if free energy is minimized at every point in time. This means that agents can minimize free energy on a moment to moment basis (through perception and action) and implicitly minimize their average free energy over their lifetime, ensuring adaptive fitness – for example avoiding surprising encounters with predators. At the same time, evolution equips agents with (adaptive) prior beliefs that define what is surprising and enables momentary minimization of free energy. In brief, the free-energy story does fix the "job relative to which" agents are assessed. That job is to minimize free-energy over time, nothing more, and nothing less. The longer you maintain a low free energy, the longer you exist. The only thing that the free-energy principle is trying to explain is how biological systems conserve themselves in the face of a changing environment. By definition, this entails a minimization of their entropy, which is simply the long-term average of surprise (under locally ergodic assumptions).

Philosopher: You are saying that given an agent (a phenotype, located in a context or niche) there will be an answer to the question "how should *this very agent* act and process information so as best to minimize surprise (free energy)" But is not part of what we need to explain the origin of that very agent, whose features and properties help determine what will count as surprise?

Physicist: I think you are missing an important feature of the story, which is that there is only one model that has the minimum free energy (for a given environment). The free-energy principle provides the mapping from the space of all possible models (which, for even one phenotype, may be uncountably large) to a single model that has the greatest chance of survival. This is the model that minimizes surprise and conserves itself. You are perhaps disappointed that this is no "super-model" that fits all environments; a model with quintessential characteristics that transcends embodiment (although environments that sustain self-organizing agents may have common characteristics that are instilled into the models that inhabit them). However, the free-energy principle does prescribe a unique (optimum) agent for each environment. I do not mean this in a hand waving way: If you specified a particular environment, you could use the free-energy principle to specify its optimum inhabitant. This is because your specification of the environment is an implicit generative model and that model will have the lowest free energy of all models.

Philosopher: What we normally mean by environment is something that many animals can share, each adapted in its different (often mutually interacting) ways. Surely the free-energy principle cannot tell us which one, among the many inhabitants of, say, a woodland niche, is optimal? I am not sure I even know what optimal can mean there. Surely they are all "optimal" in your sense, insofar as they are able, for a time, to avoid the surprising states that either constitute or lead to their own non-existence. As Theorist said, is not the "*reduce-surprise/live-longer*" hypothesis just a kind of tautology*?*

Physicist: Yes and no. The tautology we are talking about here is about surprise minimization. In other words, agents that frequent unsurprising states are in those states frequently. However, the free-energy principle is not a surprise-principle. A principle of

minimum surprise is a tautological truism – the free-energy principle explains how that truism is realized. Its explanatory power is quite substantial. For example, the free-energy bound on surprise tells us that adaptive agents must perform some sort of recognition or perceptual inference. I would challenge Information Theory (or Philosophy) to come up with a similarly simple account of why animate organisms perceive.

It would certainly be possible to rank woodland creatures in terms of their average surprise (provided one was able to measure their sensations) because this is just the entropy of their sensory states over time. Those phenotypes who maintained low entropy distributions for short periods of time would have higher average surprise and lower adaptive fitness. In other words, some creatures (models) are more optimal than others. Having said this, there is a unique optimal action and state of perceptual inference for any given creature (model). These are the behaviors and percepts that minimize free energy, given a particular creature or model.

Theorist: I took Physicist's main point to be that if you "reduce surprise," you "live longer." I do not see how a particular way of bounding surprise can be relevant to that, no matter how good it is. Is this a case where a measurement formula has been mistaken for an objective property of the world? The "reduce-surprise/live-longer" hypothesis seems to contain some remnant of the assumption that surprise is somewhere "out there," a real, objective and measurable property of the world. In fact, it is subjective and relative to the interpretation applied by the agent, i.e., always "in the eye of the beholder." The fact that we are disagreeing on this reminds me of the exchange in the original "physicist" debate (the Prolog of Eddington's Space, Time, and Gravitation 1920), concerning measurement of *length*. The physicist in that debate was drawn to the idea of length being an objective property of the world. But the relativist was more inclined to the view that length is just something we measure in different ways. As with length, so with information? This would be my contention. The present debate then seems to link nicely back to Eddington's: both can be seen as exploring difficult issues relating to exploitation of apparently objective, but in fact deeply subjective properties of the world.

Physicist: I confess I would rather have been a relativist than a physicist here! In one sense, the principal argument about the Dark-Room problem is true to relativistic sentiments: surprise is only defined in relation to a model – and a model is only good in relation to another (including the free-energy principle itself). In this sense, information and surprise are indeed in the eye of the beholder and, like length, depend upon how they are measured or modeled.

Theorist: Surely no-one could argue with that. But this seems a far cry from saying we can unify all adaptive autopoietic and self-organizing behavior under the "reduce-surprise/live-longer" imperative. Application of the free-energy principle is supposed to yield, as Physicist earlier stated, a "unique optimum agent for each environment." If the measurement of surprise (and information) itself turns out to be relative to the observer, how could this be so?

Physicist: This is simple: because surprise is "relative to" or conditioned on the observer (model) it becomes an attribute of the model. The observer (model) with the lowest average surprise, over all possible observers, is optimal for a given environment. Surprise and free energy are quantities that measure the relationship between an agent and its environment. This was the key to resolving the dark-room problem and resonates with your relativistic conclusions above.

Philosopher: I shall attempt a summation, without implying full agreement between the parties. The free-energy principle relies on free energy (bounding surprise) being defined relative to a model. We must here understand "model" in the most inclusive sense, as combining interpretive disposition, morphology, and neural architecture, and as implying a highly tuned "fit" between the active, embodied organism and the embedding environment. The dark-room scenario cannot then obtain unless staying in the dark room is itself the way to minimize surprise relative to the expectations implicitly defined by that entire, and importantly niche-reflecting, model. That means that for creatures like us, the dark room is simply not an attractor. The explanatory burden then shifts to the question how we became creatures like us in the first place. Here too, we may appeal to competition between models, again in that inclusive sense. Here too, the models (creatures) that minimize free energy will be the ones that come to populate the world. It is not obvious to Philosopher, however, that every cognitively important feature of every such model (creature) will be a direct or indirect reflection of the free-energy minimization mandate. This is the missing element that Physicist (unlike Theorist) believes is implicit in the mathematics.

Theorist: No doubt the debate will continue. Ultimately it may be resolved through accumulation of empirical evidence. This is an area where we enjoy the benefits of being able to test hypotheses using computer simulation. In due course, realistic working models will be forthcoming, at which stage this philosophical debate will rightly give way to detailed empirical evaluation of the claims being made.

Physicist: I think that is absolutely right. Furthermore, it clarifies what the free-energy principle brings to the table. Throughout this discussion we have referred repeatedly to the time-average of free energy, which is called "action" in physics. This means that the free-energy principle is nothing more than principle of least action, applied to information theory. In the same way that the principle of least action does not, in itself, describe the trajectory of a planet or the course of a river, the free-energy principle will need to be unpacked carefully in each sphere of its application. This is the real challenge ahead. For me, this will entail simulating and explaining adaptive behavior and the perceptual inference upon which that behavior rests. I include here perceptual categorization, synthesis, and learning. I include action, perception of action, and its understanding. I include attention, working memory, planning, and exploration. I even include theory of mind and (for the *philosopher*) self-awareness. If, in a few years time, we do not have neuronally plausible accounts of all these faculties that appeal to, and only to, free-energy minimization, then I will be surprised and will search for a better model!

ACKNOWLEDGMENTS

Karl J. Friston is funded by the Wellcome Trust (and is grateful to his parents for making him read Eddington, 1920 at an impressionable age).

REFERENCES

Barr, T. C. Jr., and Holsinger, J. R. (1985). Speciation in cave faunas. *Annu. Rev. Ecol. Syst.* 16, 313–337.

Buason, G., Bergfeldt, N., and Ziemke, T. (2005). Brains, bodies, and beyond: competitive co-evolution of robot controllers, morphologies, and environments. *Genet. Program. Evol. Mach.* 6, 25–51.

Conant, R. C., and Ashby, R. W. (1970). Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* 1, 89–97.

Dayan, P., Hinton, G. E., and Neal, R. (1995). The Helmholtz machine. *Neural Comput.* 7, 889–904.

Eddington, A. (1920). *Space, Time, and Gravitation. An Outline of the General Relativity Theory*. Cambridge: Cambridge University Press, 1–16.

Fiorillo, C. A. (2010). Neurocentric approach to Bayesian inference *Nat. Rev. Neurosci.* 11, 605.

Friston, K. (2010a). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–38.

Friston, K. (2010b). Some free-energy puzzles resolved: response to Thornton. *Trends Cogn. Sci. (Regul. Ed.)* 14, 54–55.

Friston, K., and Stephan, K. E. (2007). Free-energy and the brain. *Synthese* 159, 417–458.

Gregory, R. L. (1968). Perceptual illusions and brain models. *Proc. R. Soc. Lond. B. Biol. Sci.* 171, 179–196.

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behav. Brain Sci.* 27, 377–396.

Helmholtz, H. (1866/1962). "Concerning the perceptions in general," in *Treatise on Physiological Optics*, 3rd Edn, Vol. III, ed. J. Southall, trans. (New York: Dover).

Hinton, G. E., and van Camp, D. (1993). "Keeping neural networks simple by minimizing the description length of weights," in *Proceedings of COLT-93*, New York, 5–13.

Kirsh, D. (1996). Adapting the environment instead of oneself. *Adapt. Behav.* 4, 415–452.

Mumford, D. (1992). On the computational architecture of the neocortex. II. *Biol. Cybern.* 66, 241–251.

Rao, R. P., and Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* 2, 79–87.

Sella, G., and Hirsh, A. E. (2005). The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. U.S.A.* 102, 9541–9546.

Sellars, W. (1962). "Philosophy and the scientific image of man," in *Frontiers of Science and Philosophy*, ed. R. Colodny (Pittsburgh, PA: University of Pittsburgh Press), 35–78.

Thornton, C. (2010). Some puzzles relating to the free-energy principle: comment on Friston. *Trends Cogn. Sci. (Regul. Ed.)* 14, 53–54.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2005). Constructing free-energy approximations, and generalized belief propagation algorithms. *IEEE Trans. Inf. Theory* 51, 2282–2312.

Yuille, A., and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci. (Regul. Ed.)* 10, 301–308.

Ziemke, T. (2002). Introduction to the special issue on situated, and embodied cognition. *Cogn. Syst. Res.* 3, 271–274.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 14 November 2011; paper pending published: 27 January 2012; accepted: 11 April 2012; published online: 08 May 2012.

Citation: Friston K, Thornton C and Clark A (2012) Free-energy minimization and the dark-room problem. Front. Psychology ***3****:130. doi: 10.3389/fpsyg.2012.00130*

This article was submitted to Frontiers in Perception Science, a specialty of Frontiers in Psychology.